

The Role of Computation in Biodefense

The Conference on High-Speed Computing
LANL / LLNL / SNL

Salishan Lodge, Gleneden Beach, Oregon
24 April 2003

Murray Wolinsky
murray@lanl.gov

Overview

1. Biothreat 101
2. Bioinformatics 101

Examples

3.	Sequence analysis:	mpiBLAST	Feng
4.	Detection:	KPATH	Slezak
5.	Protein structure:	ROSETTA	Strauss
6.	Real-time epidemiology:	EpiSIMS	Eubank
7.	Forensics:	VESPA	Myers, Korber

8. Needs

System level analytical capabilities
Enhanced phylogenetic algorithms
Novel countermeasures

1. Biothreat 101: Historical Perspective

- 1932-45 Japanese conduct biological experiments in Manchuria
- 1972 Biological Weapons Convention
- 1980's Iraq employs chemical weapons against Iranians and Kurds
- 1995 Aum Shinrikyo releases home-brewed Sarin nerve gas in a Tokyo subway. 12 deaths and 5500 hospitalized
- 1996 Evidence that Iraq had produced and was prepared to use 19,000 pounds of botulinum toxin and 8,500 pounds of anthrax in the Gulf War
- 1997 U.A. Army announces successful anthrax vaccine. Russians publish genetically engineered strains of vaccine-resistant anthrax
- 1998 US government releases data about the extensive bioweapons program in the FSU – *B. anthracis*, *Y. pestis*, Ebolapox
- 2001 Anthrax powder mailed through US postal system in wake September 11

Historical Perspective

- In the 1960's the US Surgeon General told Congress that infectious diseases had been conquered.
- Tuberculosis is spreading world-wide
 - 8-10 million new cases each year
(10% resistant to several of the front-line drugs)
- Resurgence of several infectious diseases, e.g.,
 - 300-500 Million new cases of malaria each year
 - 50-100 Million new cases of Dengue fever each year
- HIV emerged in the early 80's
 - Levelled off in the US
 - 30 to 50 Million cases world wide by the end of the century
- 90 Thousand deaths/year in the US in a health care setting (from nosocomial infections) – 42 isolates from local hospital resistant to vancomycin; increased from 4 to 42 in 3 years
- SARS

Scale of the biotreat

<i>Type</i>	<i>Fatalities</i>	<i>Likelihood</i>
Efficient biological attack	1,000,000	extremely low
Atomic bomb detonated in major city	100,000	very low
Attack on nuclear or toxic chemical plant	10,000	very low
Inefficient biological or chemical attack in a skyscraper	1,000	low

Office of Technology Assessment, *Proliferation of Weapons of Mass Destruction: Assessing the Risks* (U. S. Congress, 1993)

Biothreat 101: an idiosyncratic taxonomy

Naturally-occurring					
Familiar					
Unfamiliar (e.g., SARS)		public health activities			
Anthropogenic					
Accidental					
Intentional					
Conventional	intel	targeted			advanced
Engineered					
	Prevention	Detection	Intervention	Treatment	Forensics
		3,4,5	6	5	7

2. Bioinformatics 101

Biology

important organisms: e.g., humans, plants, bacteria, viruses

Two types of important molecules

DNA (RNA) -- string of 4 nucleotides -- double helix

proteins -- string of 20 amino acids -- complex structures

Central dogma of bioinformatics

sequence ---> structure ---> function

PCR (primers)

Sequencing (contigs, assembly, finishing)

Annotation (gene prediction, alignment, BLAST, Genbank)

Databases

Protein structure prediction

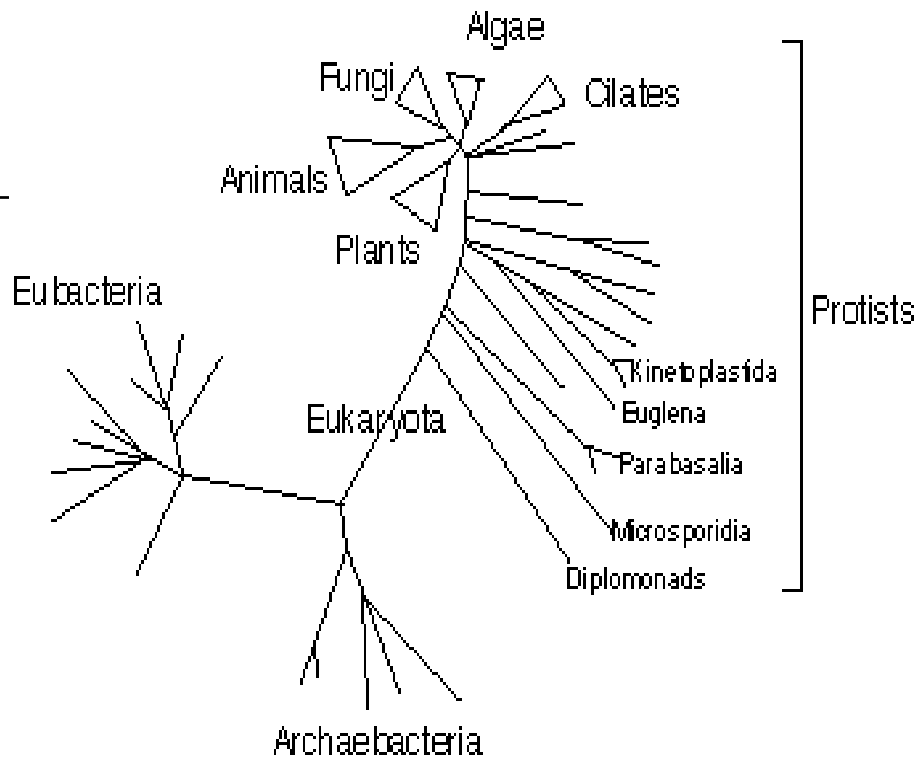
Perl -- “the language that saved bioinformatics”



2. Bioinformatics 101

Biology

important organisms: e.g., humans, plants, bacteria, viruses



genotype vs. phenotype

polymorphisms

neutral mutations

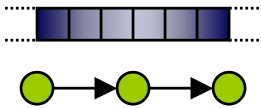
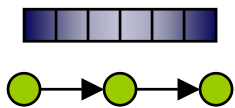


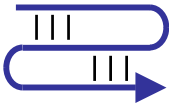
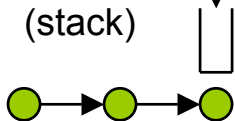




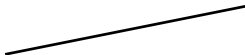

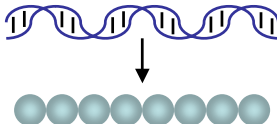
single point mutation

genetic recombination

algorithms for inverting evolution

Larry Hunter, *Molecular Biology
for Computer Scientists*

The Chomsky Hierarchy

Language	Automaton	Grammar	Recognition	Dependency	Biology
Recursively Enumerable Languages	Turing Machine 	Unrestricted $Baa \rightarrow A$	Undecidable	Arbitrary	Unknown
Context-Sensitive Languages	Linear-Bounded 	Context-Sensitive $At \rightarrow aA$	NP-Complete 	Crossing 	Pseudoknots, etc. 
Context-Free Languages	Pushdown (stack) 	Context-Free $S \rightarrow gSc$	Polynomial 	Nested 	Orthodox 2° Structure 
Regular Languages	Finite-State Machine 	Regular $A \rightarrow cA$	Linear 	Strictly Local 	Central Dogma 



Some Bioinformatic Applications in Biothreat Reduction

- Signature development
 - Unique to specific organism *phylogenetics*
 - Tied to mechanism of pathogenesis *annotation/analysis*
- Attribution *phylogenetics/geography*
- Novel countermeasures *structure prediction*
- Data access and exchange *XML*

Bioinformatics for bio-threat reduction emerged at Los Alamos from early work in health related areas

1982: GenBank

1986: HIV Sequence Database

1994: Papilloma Virus Database

1996: Influenza Database

1998: Sexually Transmitted Diseases Database

1999: CBNP Databases

2000: NIH Oral Microbial Pathogen Database

2002: USAMRIID Toxin/Virulence Database

3. mpiBLAST: Delivering Super-Linear Speed-Up with an Open-Source Parallelization of BLAST

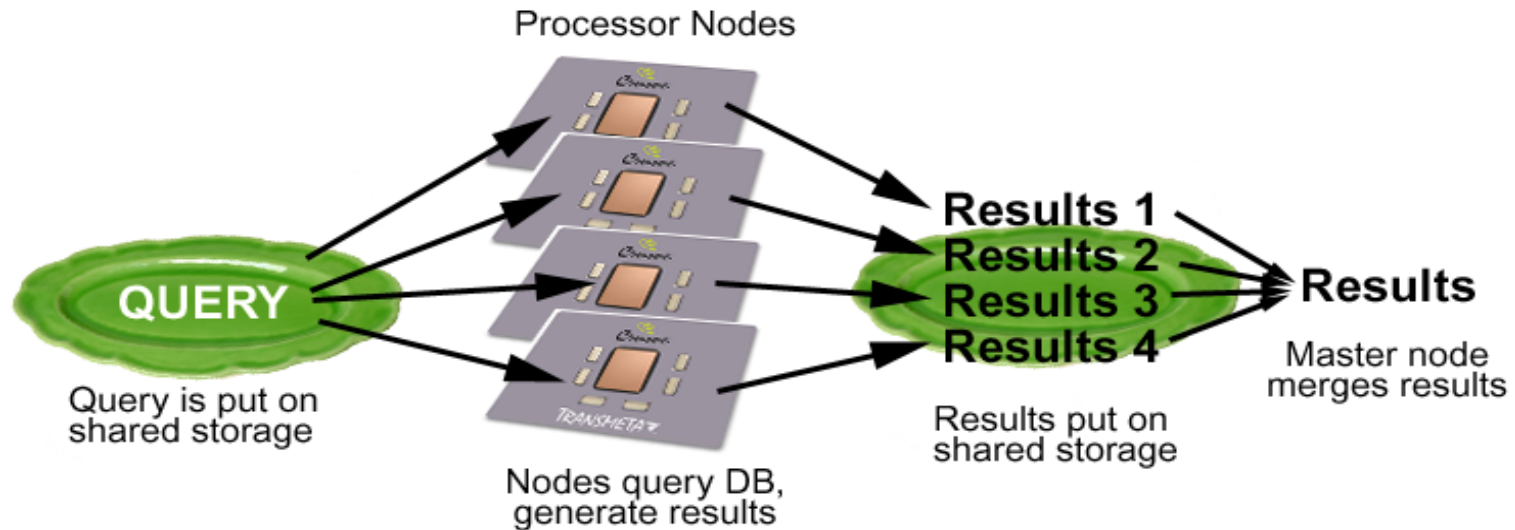
Wu-chun (Wu) Feng
feng@lanl.gov



Parallelizing BLAST

- Multithreading
 - Implemented in NCBI's BLAST.
- Query Segmentation
 - Divides a query into sub-queries and each sub-query is searched against a copy of the entire database on each node.
 - Many implementations exist.
- Database Segmentation
 - Fragments the database into smaller pieces where each piece fits entirely in memory. Each cluster node searches on one fragment of the database.
 - Only known open-source implementation: mpiBLAST.

Database Segmentation

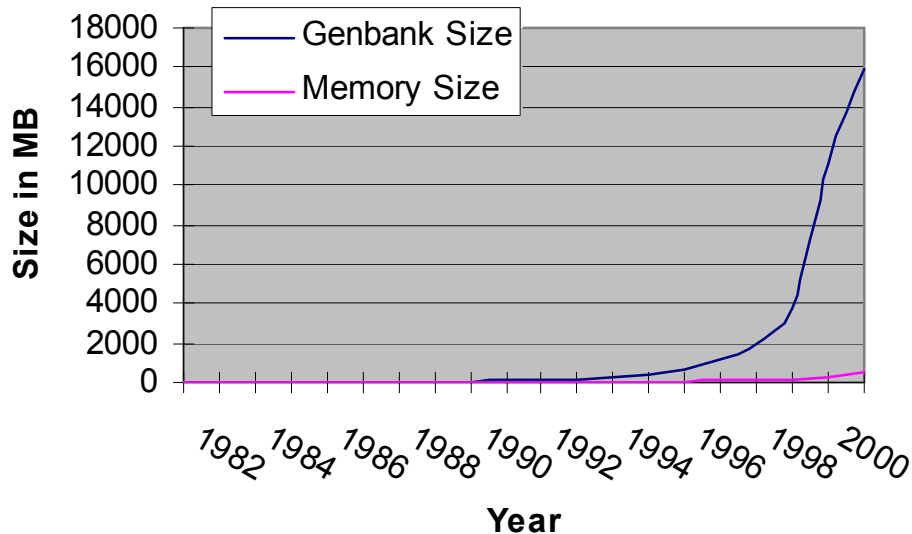


- Since database distribution occurs only once, its cost is amortized across all subsequent queries to the database. The database changes infrequently and is periodically appended with new sequences.

Enormous Sequence Databases

Size in MB	DB name	Description
5700	nt	non-redundant nucleotide DB
2200	Human EST	Human expressed sequence tag DB
1100	Mouse EST	Human expressed sequence tag DB
510	nr	non-redundant amino acid DB

Growth of Genbank vs. Memory Size



*Growth Trend:
Database Size
vs. Memory Size*

mpiBLAST Performance

BLAST Run Time for a 300kb Query against nt :

Nodes	Runtime (s)	Speedup over 1 node
1	80774.93	1.00
4	8751.97	9.23
8	4547.83	17.76
16	2436.60	33.15
32	1349.92	59.84
64	850.75	94.95
128	473.79	170.49

Reduces search time ...

From over 1346 minutes (22.4 hours) to under 8 minutes!

Current Success of mpiBLAST

- Publications (Peer-Reviewed)
 - “The Design, Implementation, and Evaluation of mpiBLAST,” *ClusterWorld 2003*, **Best Paper Award**, Jun. 2003.
 - “mpiBLAST: Parallelization of BLAST for Computational Clusters” (poster), *SC 2002: High-Performance Networking and Computing Conference*, Nov. 2002.
 - A. Darling and W. Feng, “BLASTing Off with Green Destiny” (poster), *IEEE Computer Society Bioinformatics Conference (CSB'02)*, Aug. 2002.
- Recent Media Coverage
 - “LANL Researchers Outfit the 'Toyota Camry' of Supercomputing for Bioinformatics Tasks,” *BioInform/GenomeWeb*, Feb. 2003.
- Downloads
 - Nearly 400 institutional downloads in only two weeks time.



Future Directions

- Making mpiBLAST Even Faster
 - Automate database fragmentation.
 - Couple query segmentation with database segmentation.
 - Replication of sub-queries in heterogeneous systems.
 - Re-work “mpirun” to more efficiently distribute queries and databases.
- Making mpiBLAST More Robust
 - Automate the migration (or replication) of queries and/or databases to other computing nodes.
- Making mpiBLAST More Manageable
 - Create a transparent environment for end users by presenting the cluster computer as a single machine.

4. KPATH: Tom Slezak, LLNL

- **Scaling Infrastructure**
- **Pipeline Automation**
- **Protein Structure Prediction & Analysis**
- **New Algorithms**
- **Information Integration**
- **Minimizing cost of system evolution**

TCACTCCGGC CGACAAAAGC GACAAAGGTT TTGTTCTTGG TCACTCCATA
TCAC TCCGGC CGACAAAAGC GACAAAGGTT TTGTTCTTGG TCACTCCATA
TTACTCCAGC TGACAAAAGC GACAAAGGTT TTGTTCTTGG TCACTCCATT
TTACTCCAGC TGACAAAAGC GACAAAGGTT TTGTTCTTGG TCACTCCATT
TCACTCCGGC CGACAAAAGC GACAAAGGTT TTGTTCTTGG TCACTCCATA
TTACTCCAGC TGACAAAAGC GACAAAGGTT TTGTTCTTGG TCACTCCATT

[illegible]

Foot and Mouth Disease Signature Development

gtggt.gc.ag.ga..a.ga..tgga.tt.gaggc.ct.aagcc.cactt.aa.tc.cttgg.ca.ac.at.
ac.cc.gc.GACAAAAG**CGACAAAGGTTTGTCTTGG**
TCA.tccat.ac.ga.gtcactttcctcaaag.cacttcc..at.ga.ta.gg.actgggtttta.a
aacctgtgatggc.tc.aagaccct.gaggc.ATCCTCTC**CTTTGCACGC**
CGTGGGACCAT.caggagaagttga..tccgtggcaggactcgc.gtcca
ctc.ggacc.ga.ga.taccggcg.ctctttgagcc.tt.ca.gg.ctctt.gagat.cc.AGCTA
CAGAT**CACTTTACCTGCG.TGGGT**GAACGCCGTGTG
CGG.gacg...aa

DNA signatures must hit all targets and exclude all non-targets



We have greatly scaled our local computing infrastructure

- **Over \$900K FY02 supplemental funding provided**
 - **24 CPU, 48 GB Sun compute server**
 - **8 CPU, 32 GB Sun Oracle DB server**
 - **3 TB RAID storage server**
 - **GigE to connect all major machines**
- **Investment by DOE sponsor based on success of DNA signatures developed by our team in post-9/11 action on the BASIS program**



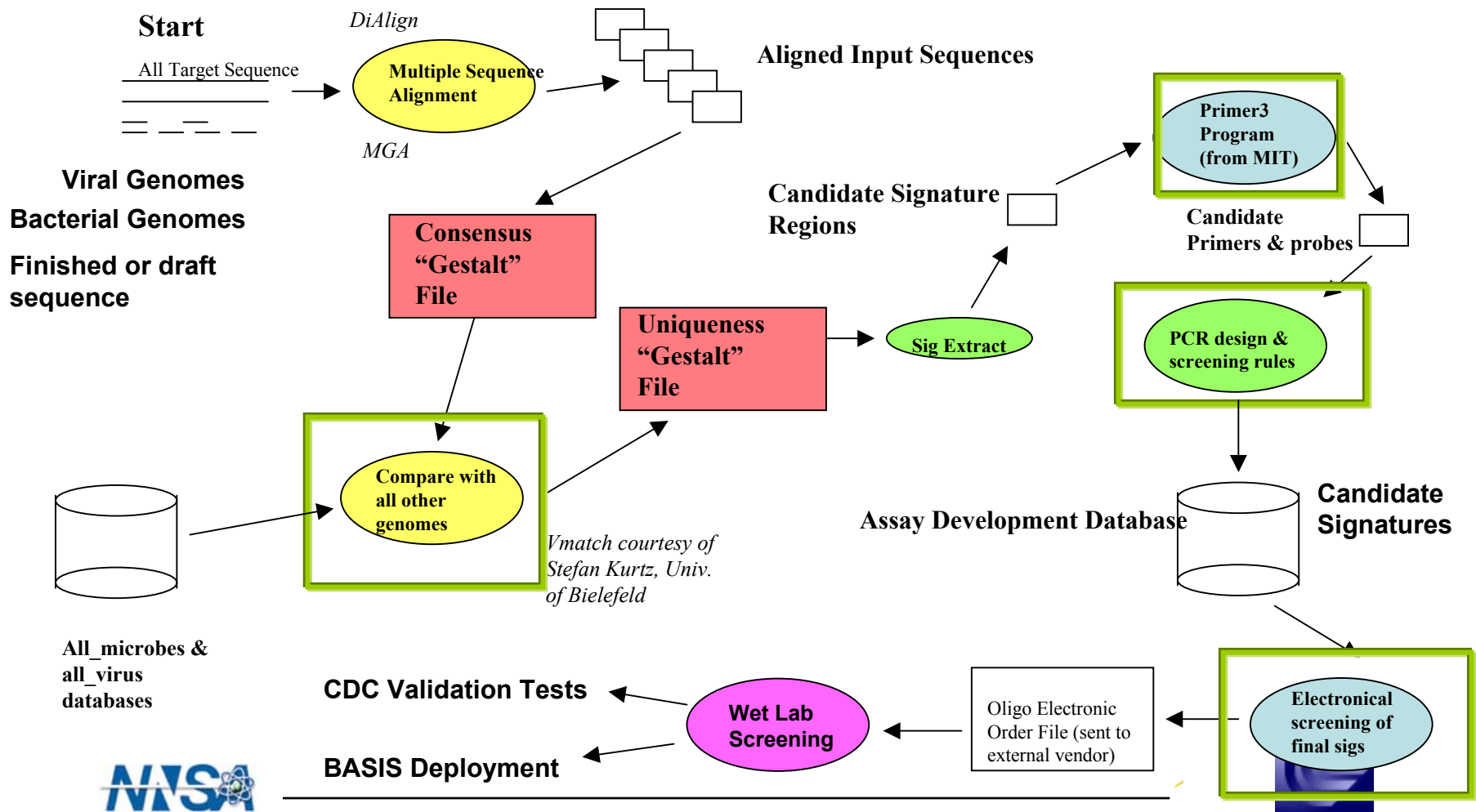
We have the only fully-automated DNA signature pipeline

- We let the genome itself tell us what is unique
 - Prior approaches focused on specific genes of interest only
- Appropriate algorithms for efficient processing
 - Multiple genome alignment
 - Suffix-array sub-string comparison
- Rigorous selection and *in silico* testing of candidate signatures; automatic evaluation of candidates as new sequence data acquired
- Results depend on quantity and quality of target and near-neighbor genomic sequence



We now run all key steps of our pipeline in parallel

We can process any pathogen in less than two hours



Our work is being published in multiple forums

- IEEE invited paper, “Rapid Development of Nucleic Acid Diagnostics, published November, 2002
- 2 Briefings in Bioinformatics invited papers accepted for June, 2003 publication:
 - An Applications-Focused Review of Comparative Genomics Tools: Capabilities, Limitations, and Future Challenges
 - Comparative Genomics Tools Applied to Bioterrorism Defense
- “Limitations of TaqMan PCR for detecting divergent viral pathogens” accepted for publication in Journal of Clinical Microbiology
- Invited book chapter in press for FBI/DOE-sponsored volume on Microbial Forensics:
 - Bioinformatics Methods for Microbial Detection and Forensic Diagnostic Design

We received one of 2 LLNL 2002 Science & Technology awards for this work

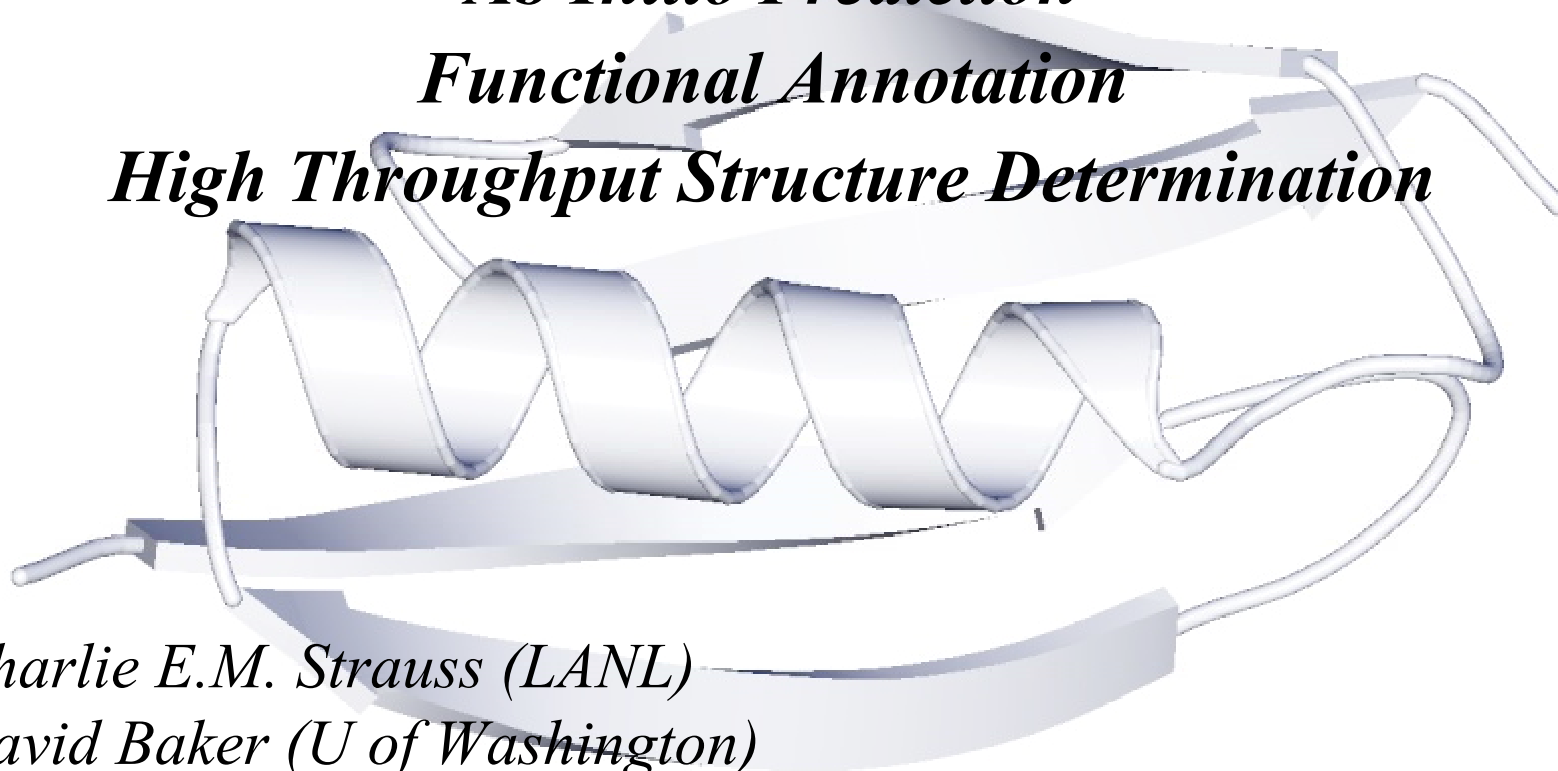


5. ROSETTA: De Novo Structure Prediction

Ab Initio Prediction

Functional Annotation

High Throughput Structure Determination



Charlie E.M. Strauss (LANL)

David Baker (U of Washington)

Richard Bonneau (Institute for Structural Biology)

Carol Rohl (UC Santa Cruz)

Structure Modeling Paradigms

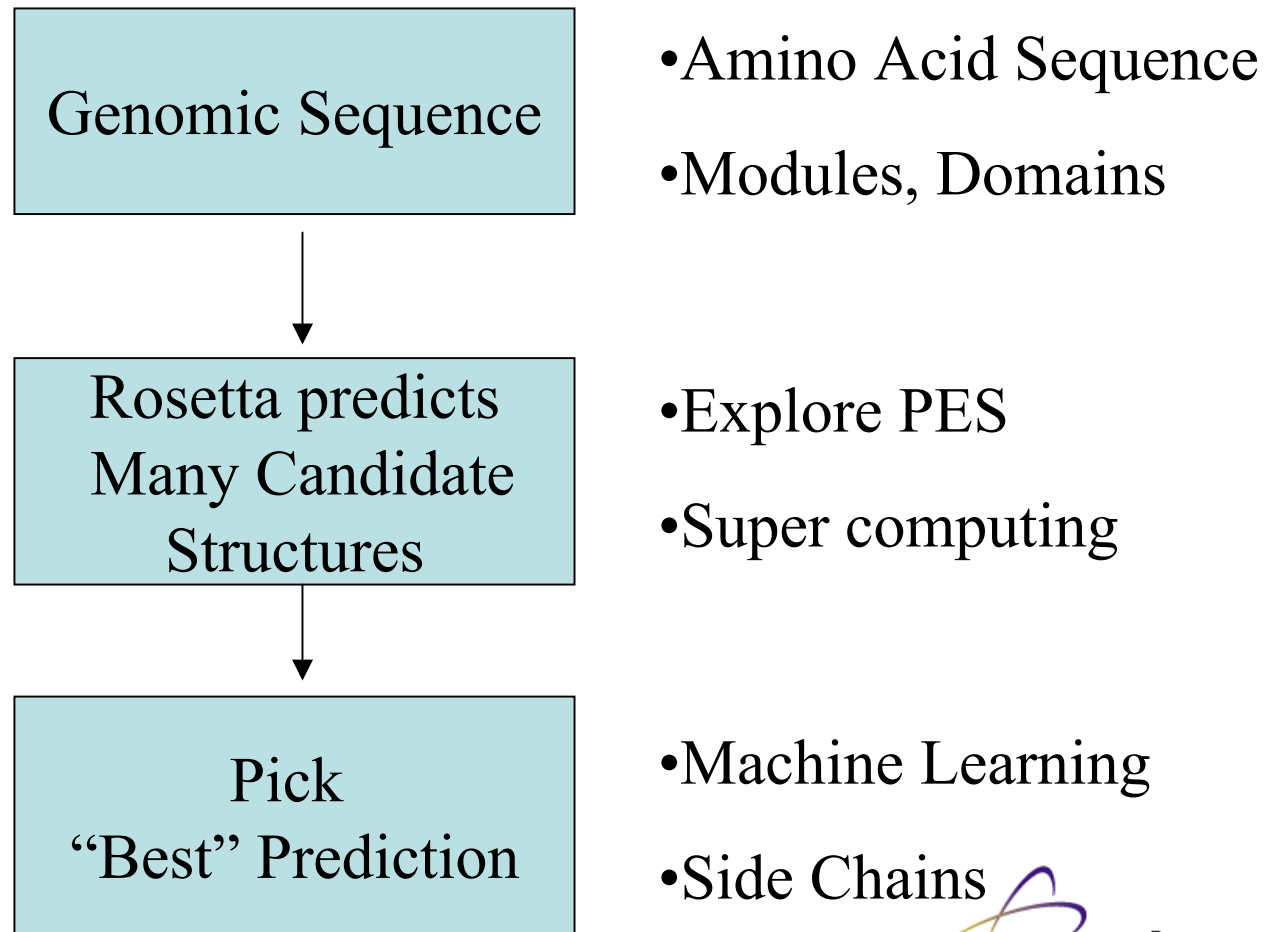
Potential Energy Surface

- **Physics-Based**
 - Approximate electromagnetic and chemical forces, energies
 - Good dynamics
- **Statistics-Based**
 - Pseudo energies based on frequencies with which inter residue relationships occur.
 - Heuristics
 - Good structures
 - Better PES for optimization.

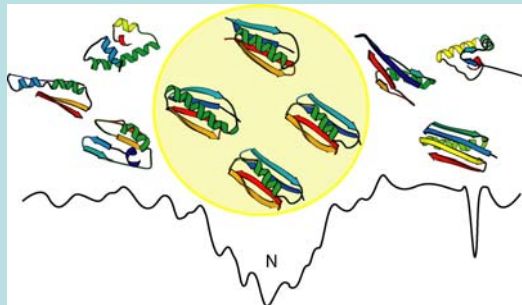
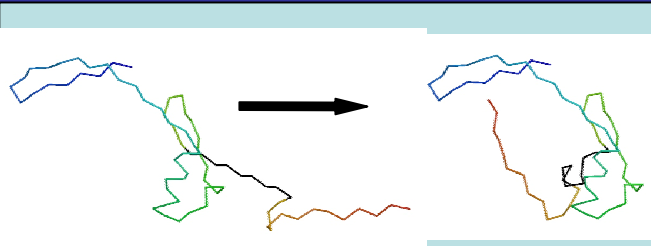
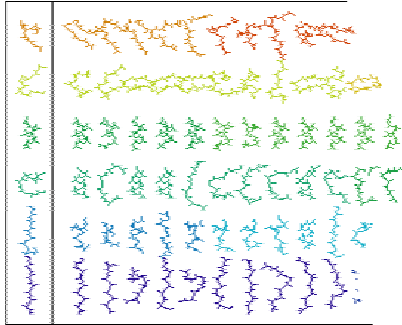
Search and Optimization

- **Comparative modeling**
 - Structure must be in database
 - Small search space
 - Large structures okay
- *Ab Initio*
 - Novel structures
 - Design
 - Conformational changes
 - Loop modeling
 - Large search space
 - Large structures (>200 aa) difficult

Ab Initio Structure Prediction



Structure Prediction with Rosetta



1. Select fragments consistent with local sequence preferences
2. Assemble fragments into models with native-like global properties

Potential Terms:

environment (solvation)
pairwise (electrostatics)

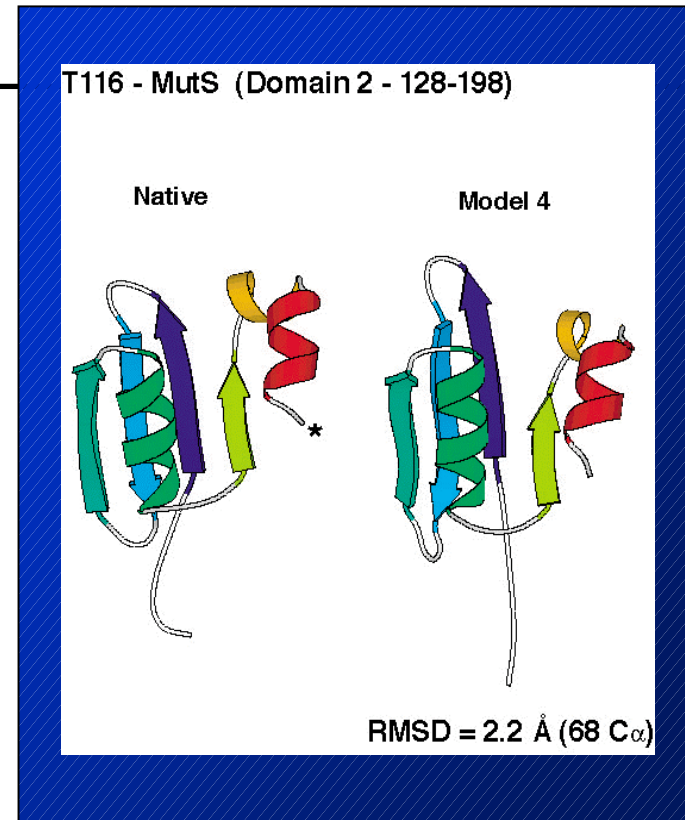
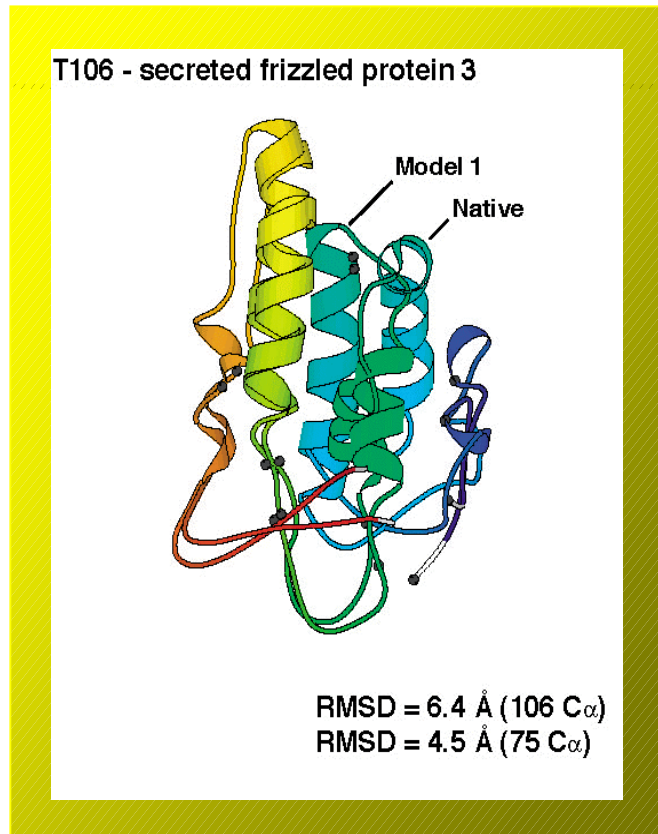
C β density
steric overlap
strand pairing
radius of gyration

Backbones with Unified Atom Sidechains

3. Identify the best model from the population of decoys

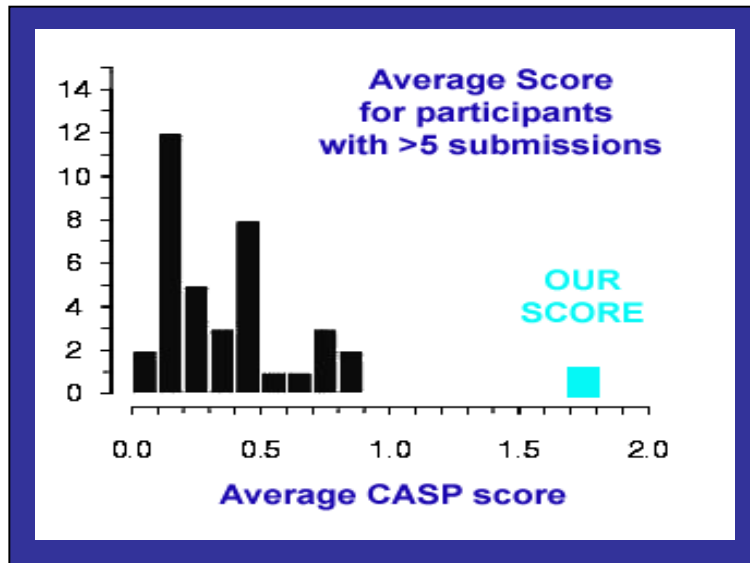
Add full atom sidechains and relax

Blind *ab initio* Predictions



- Fold Recognition
- Comparative Modeling
- Novel Fold
- Homology Modeling**
- Ab Initio* Modeling**
- Docking
- Loop Modeling

CASP 4 *Ab Initio* Summary

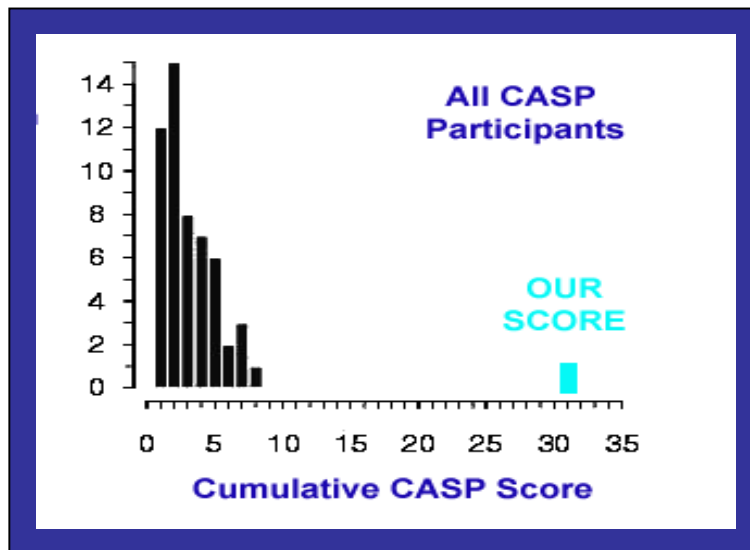


- 18 Newly solved structures predicted *prior* to publication of structure.

- True *Ab Initio* targets.

- None could be recognized by sequence similarity

- None of these even had close structural homologs.



Independently assessed scoring:

2="Well Above Average", 1="okay", 0="lousy"

Parting Thoughts and Pot Shots

- 2002 state of the art

Pure *Ab initio*: <150Residues

- Roughly 7Å rms backbone over 100 Amino acids is threshold for Scop superfamily assignment.
- Ab initio* methods can outperform “comparative modeling”
- Paradoxically, as the PDB grows, *ab initio* becomes more useful not less.
 - Loop Modeling, Functional Annotation, Design, Motifs.
 - Assist Experimental Methods

Acknowledgements

Charlie E. M. Strauss (Los Alamos National Lab)

David Baker

Richard Bonneau (Stromix)

Carol Rohl

Peter Bowers (Protein Pathways)

Dylan Chivian

Ingo Ruczinski (Johns Hopkins University)

Kim Simons (Harvard University)

Jerry Tsai (Texas A&M University)

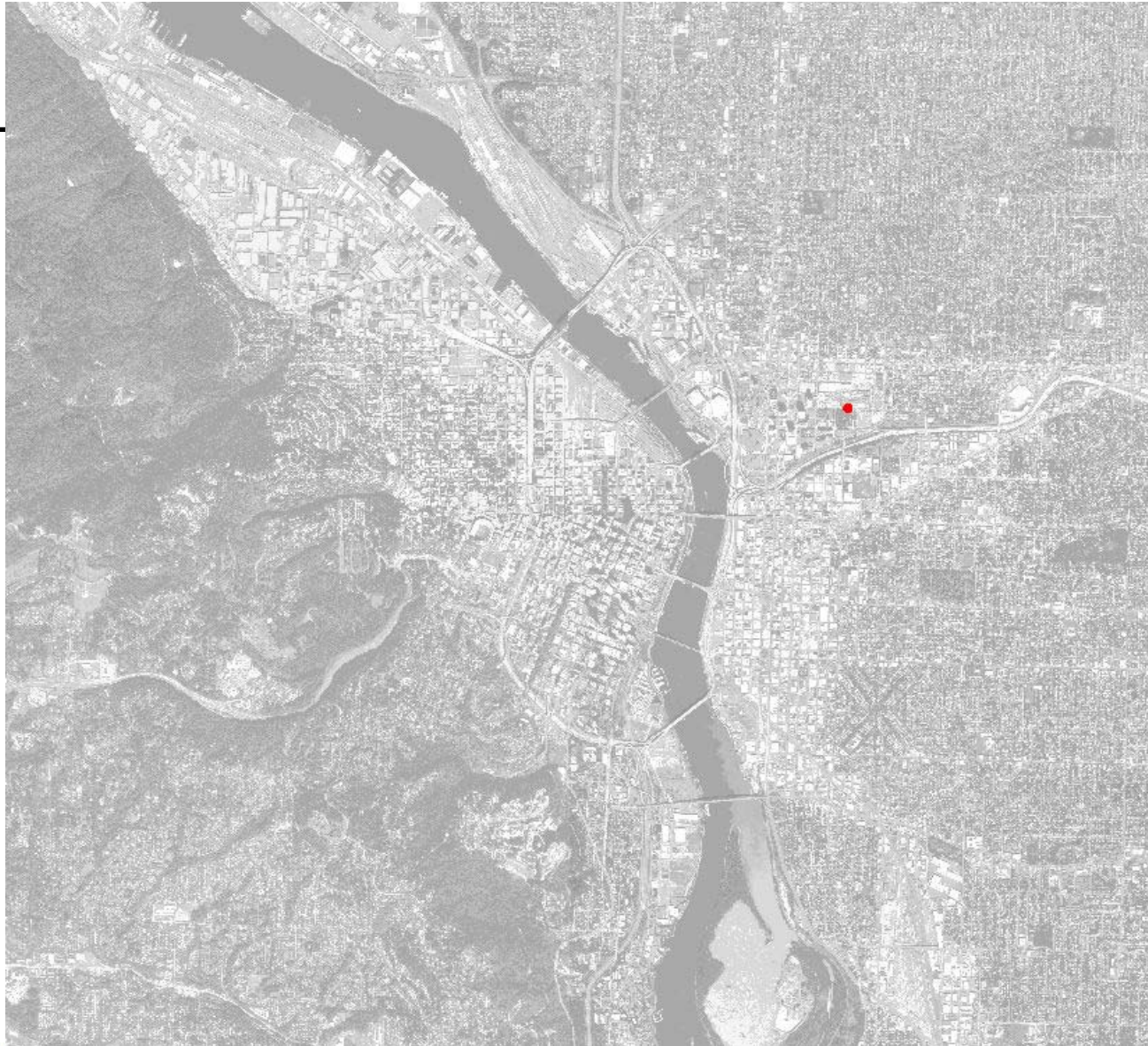
HHMI, NIH, DOE

6. EpiSIMS, Stephen Eubank

A new approach to epidemiology for decision support

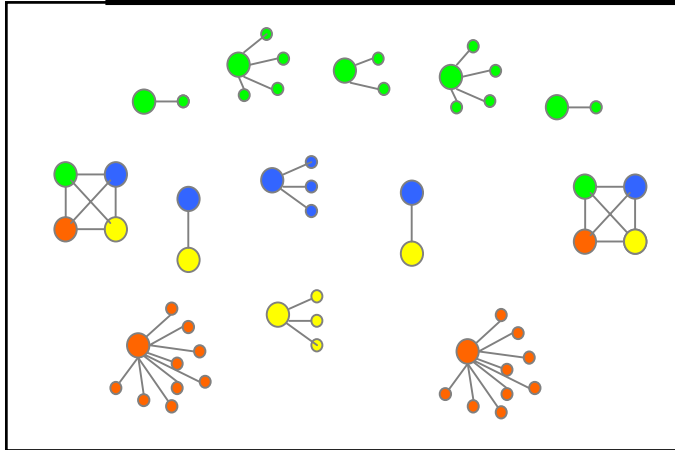
- simulate large populations at the level of individuals
- make possible analysis of interactions among
 - human activity patterns
 - disease parameters
 - targeted mitigation strategies
- useful for
 - policy assessment
 - testing infrastructure changes
 - gaming
 - real time crisis management

Day 1: release at red location

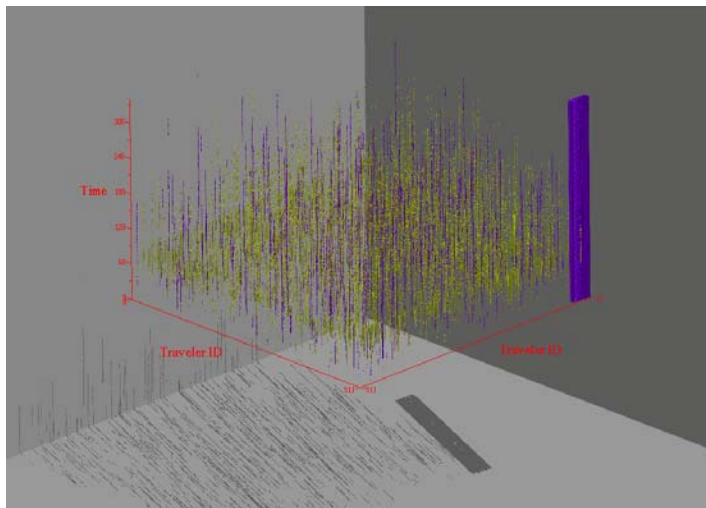
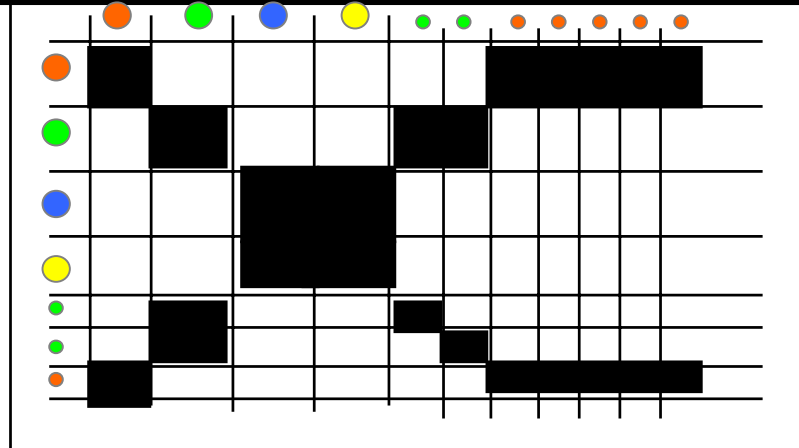


Individuals' contacts determine spread

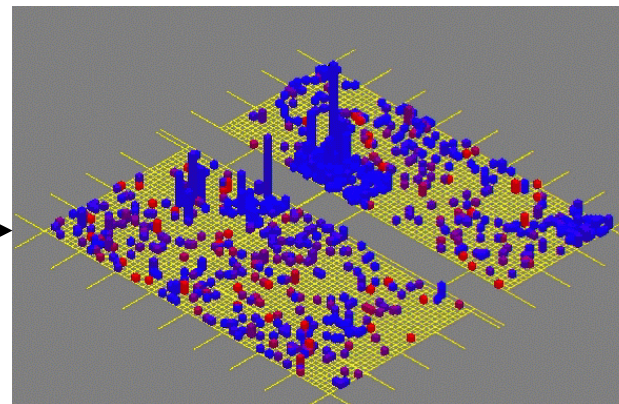
Family's activities



Contact matrix for entire population

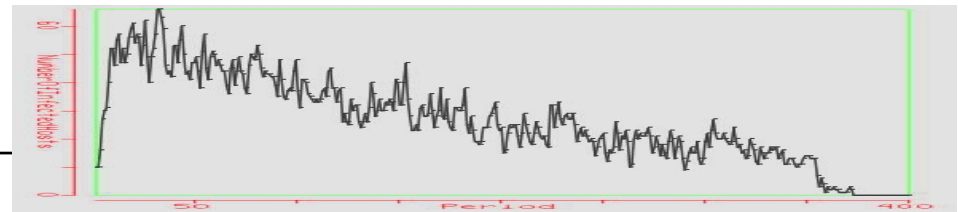


Time dependent contacts



Epidemic snapshot

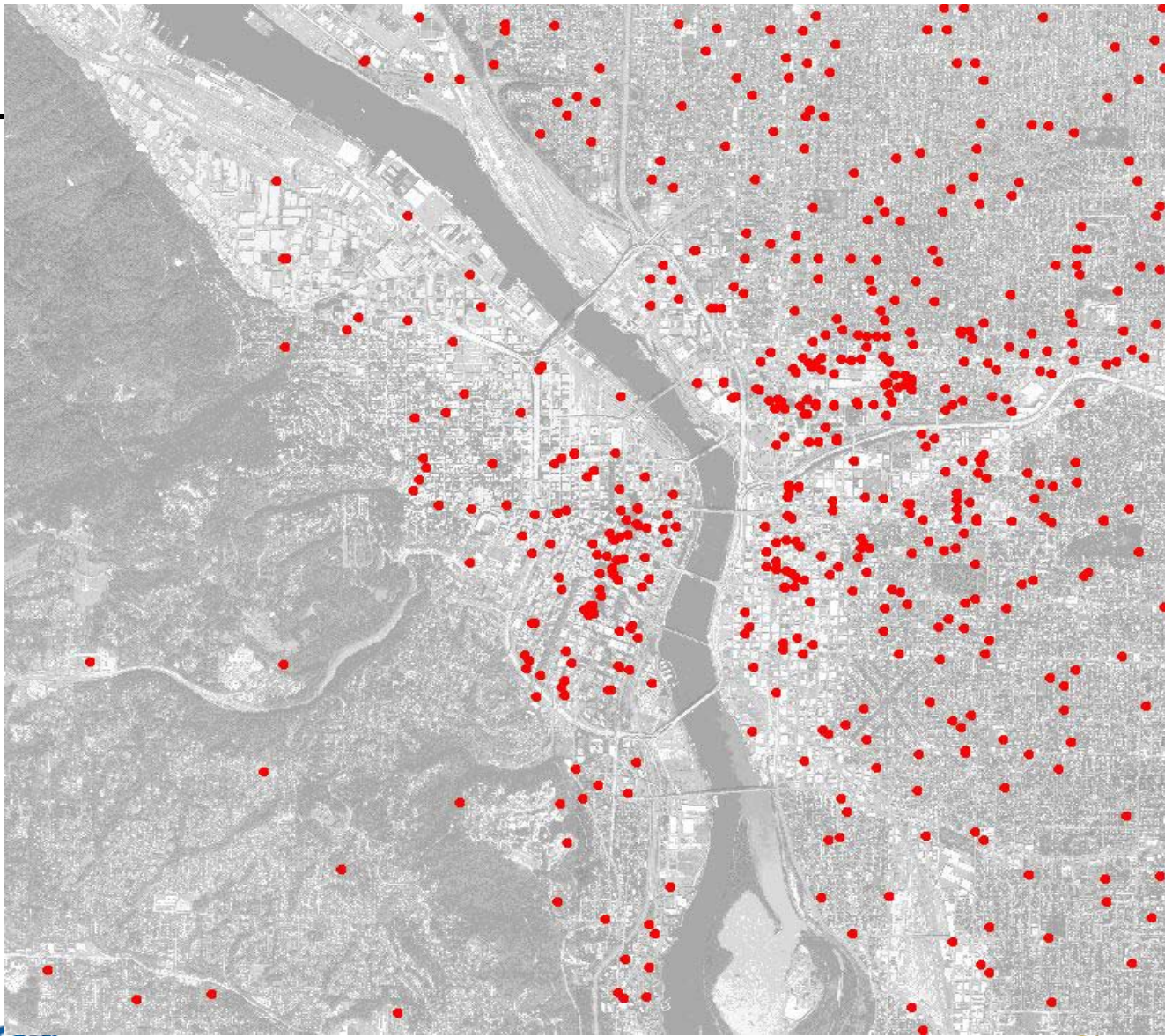
Epidemic curve



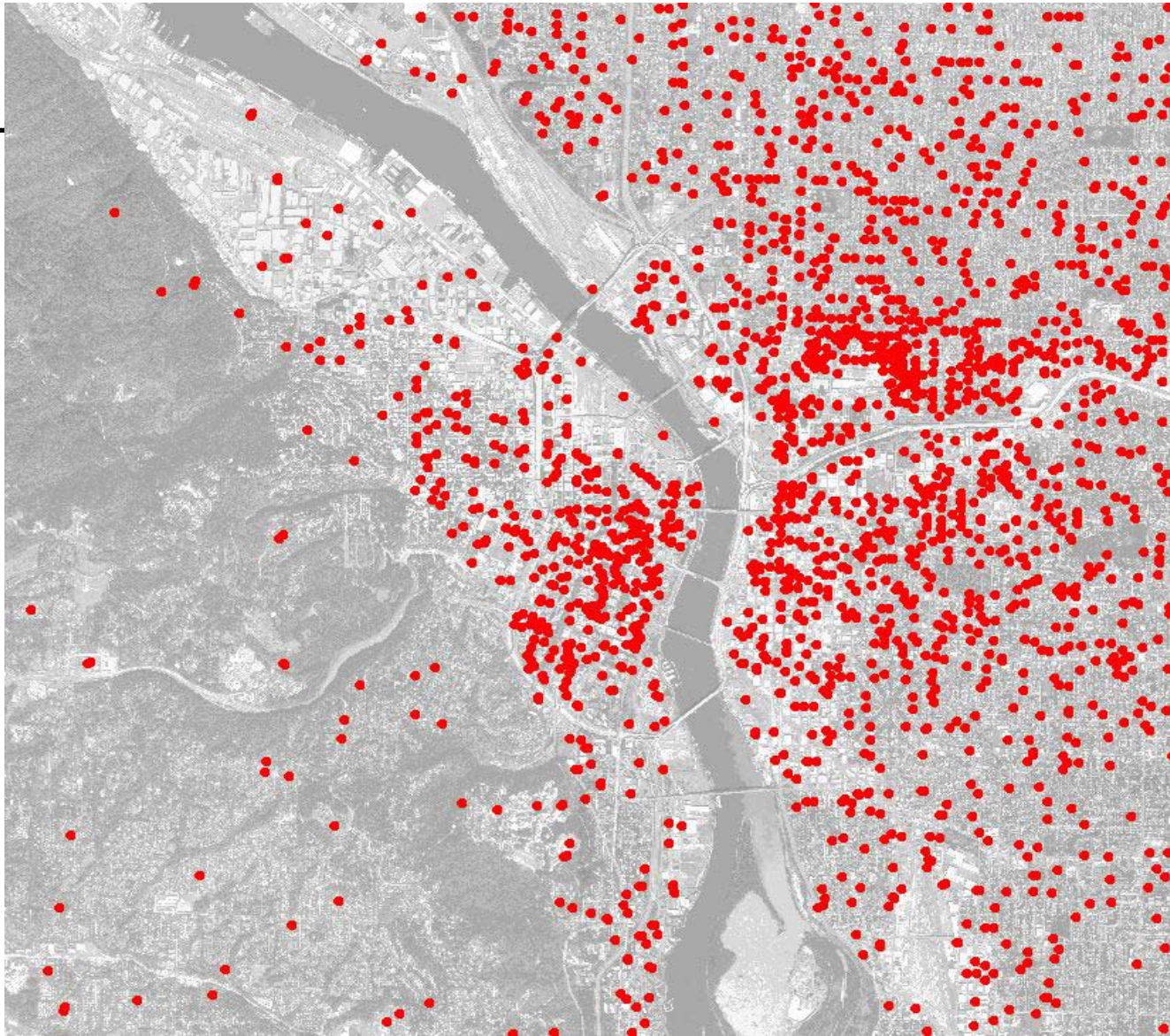
Day 2: locations w/ infected people



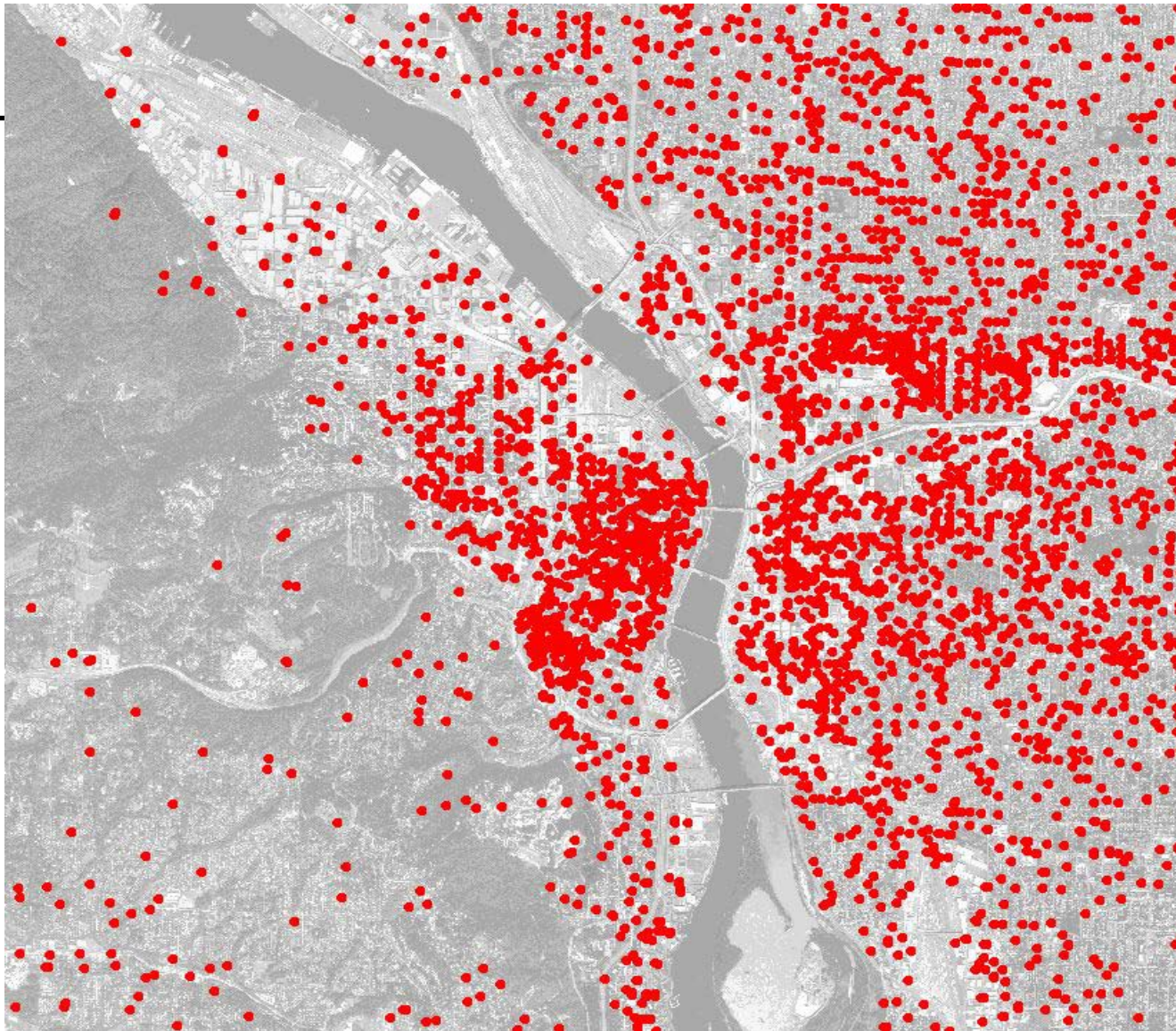
Day 4: cont'd infection from initial event



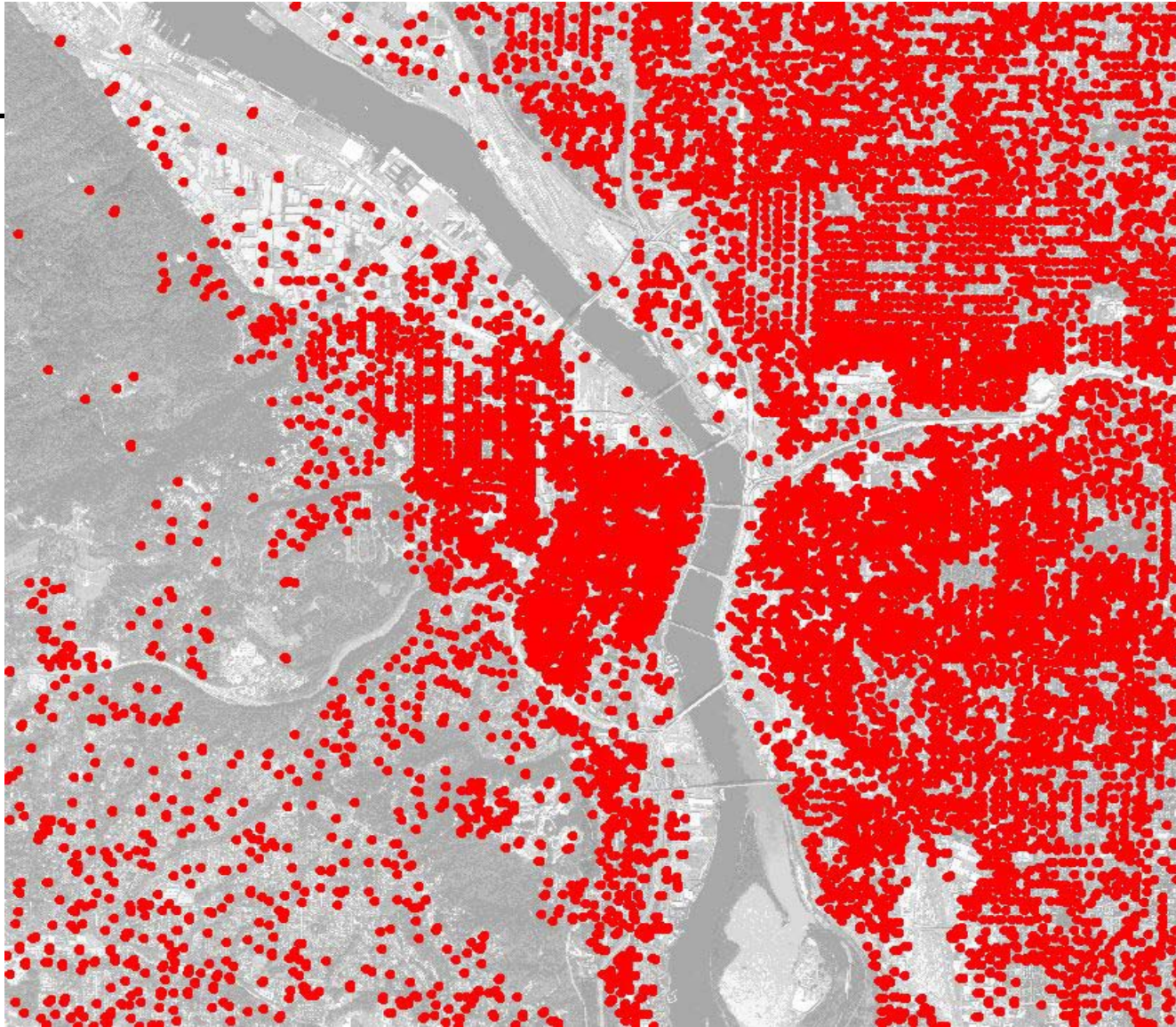
Day 8: secondary infections begin



Day 14: more secondary infections



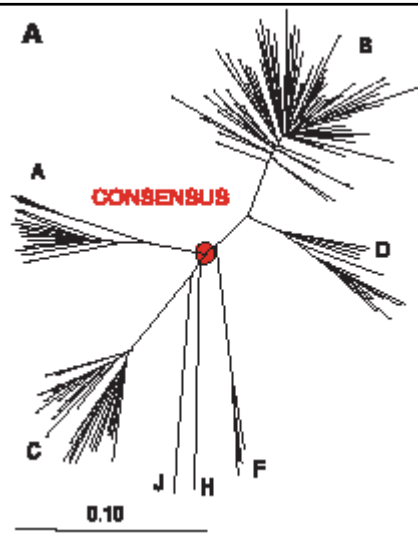
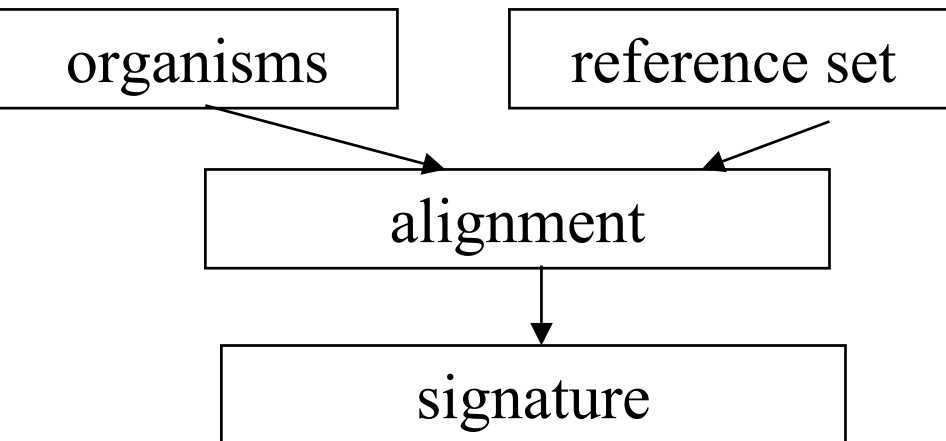
Day 17



EpiSims

- **Simulation-based decision system**
 - **scenario definition**
 - **simulation provides consequences**
 - **expressed via organizing principles: critical pathways**
 - **analysis suggests appropriate interventions: sever pathways**
 - **intervention definition**
 - **hypothetical, not necessarily achievable**
 - **generic or tailored to scenario**
 - **analysis of consequences for decision support**

7. VESPA, Myers and Korber; ML phylogeny, Korber



Viral Epidemiology Signature Pattern Analysis (VESPA)

Korber B and Myers G: Signature pattern analysis: a method for assessing viral sequence relatedness *AIDS Res. Human Retroviruses* 8(9): 1549-1560 (1992).

B. Korber et al, Timing the Ancestor of the HIV-1 Pandemic Strain, *Science*, 288, 9 June 2000, pp. 1789-1796.

8 Needs

Better query capabilities

- Significance of observed patterns
- Applications to virulence; pathogenicity islands

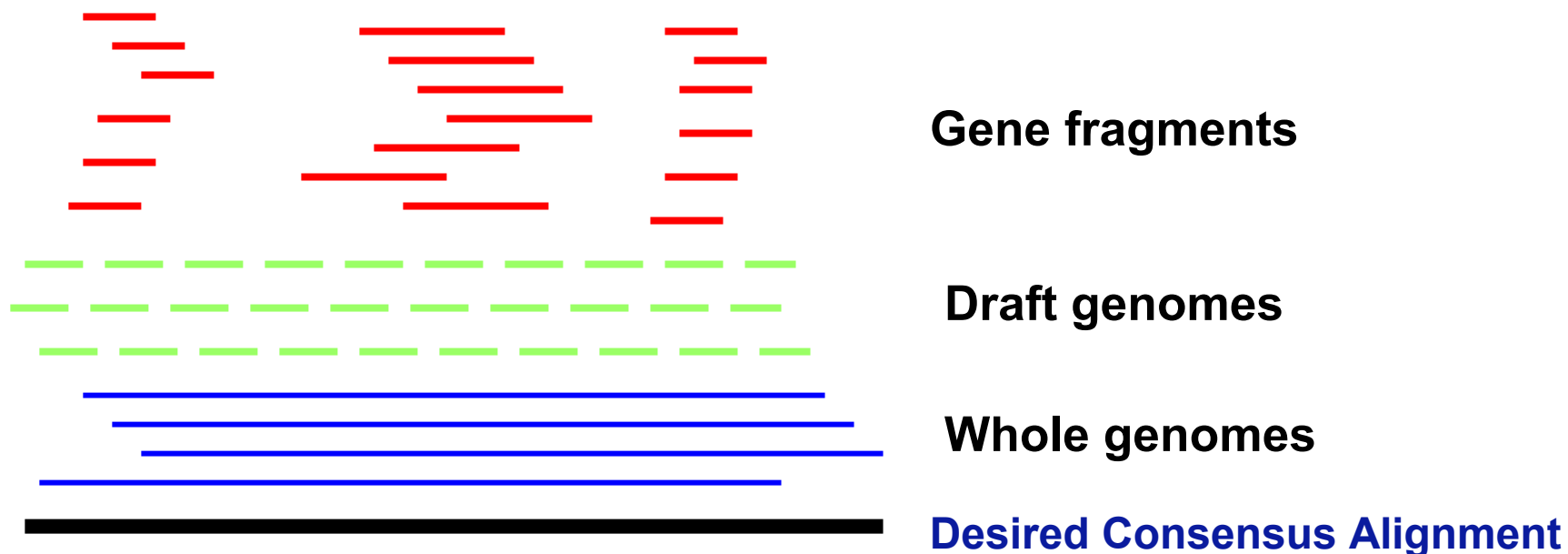


Correia repeats in *Neisseria meningitidis*

- Searches incorporating annotation
- Searches across genomes

Need new alignment algorithms for available
pathogen sequences

**Need to align finished genomes with draft
genomes and gene-fragment sequences**



Needs

Better visualization tools

Phylogenetic techniques incorporating recombination

Methods to detect engineering

Literature mining capabilities

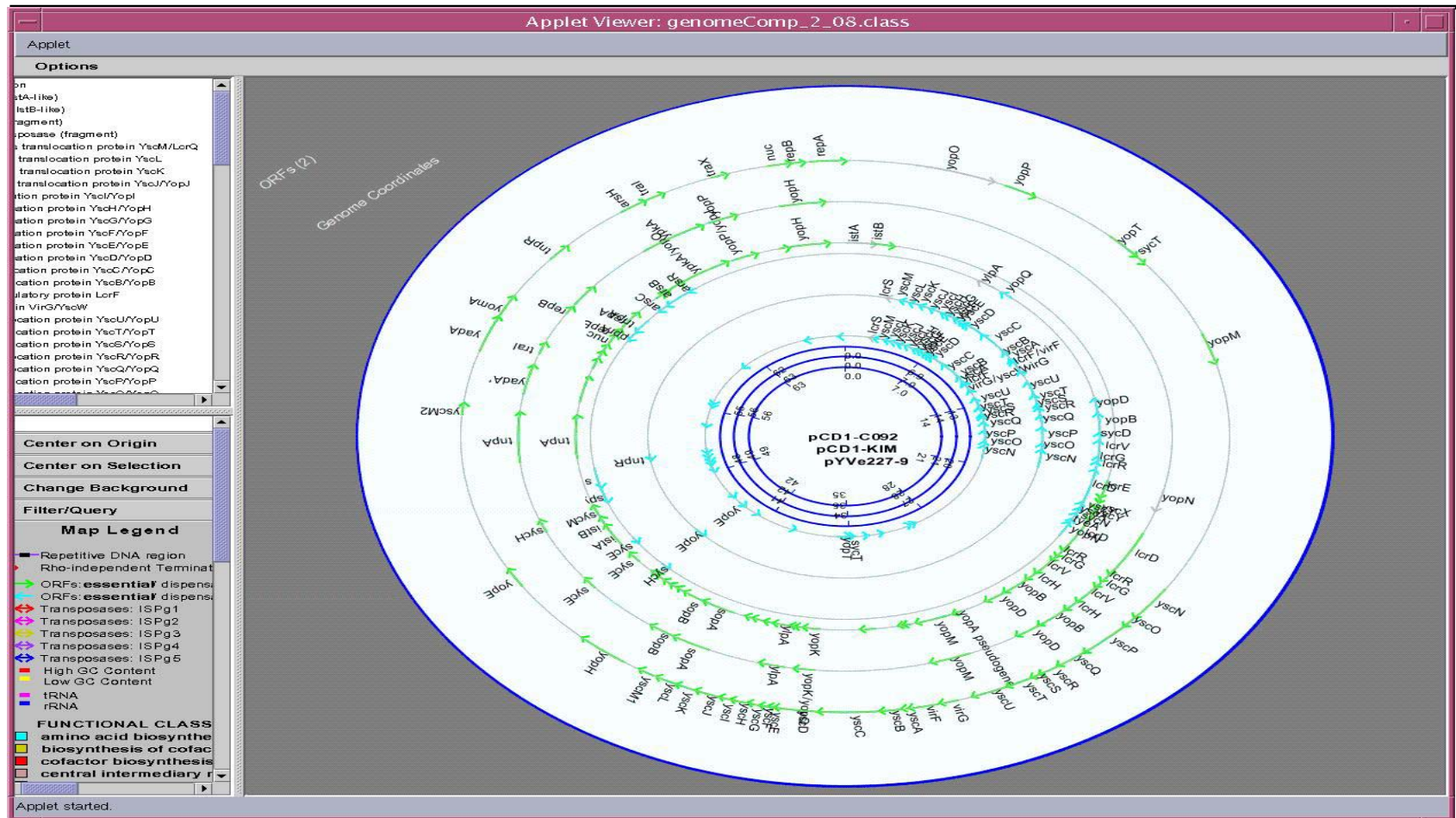
Tools to help unravel mechanisms of pathogenicity

Tools to help design improved and novel vaccines,
drugs and other countermeasures

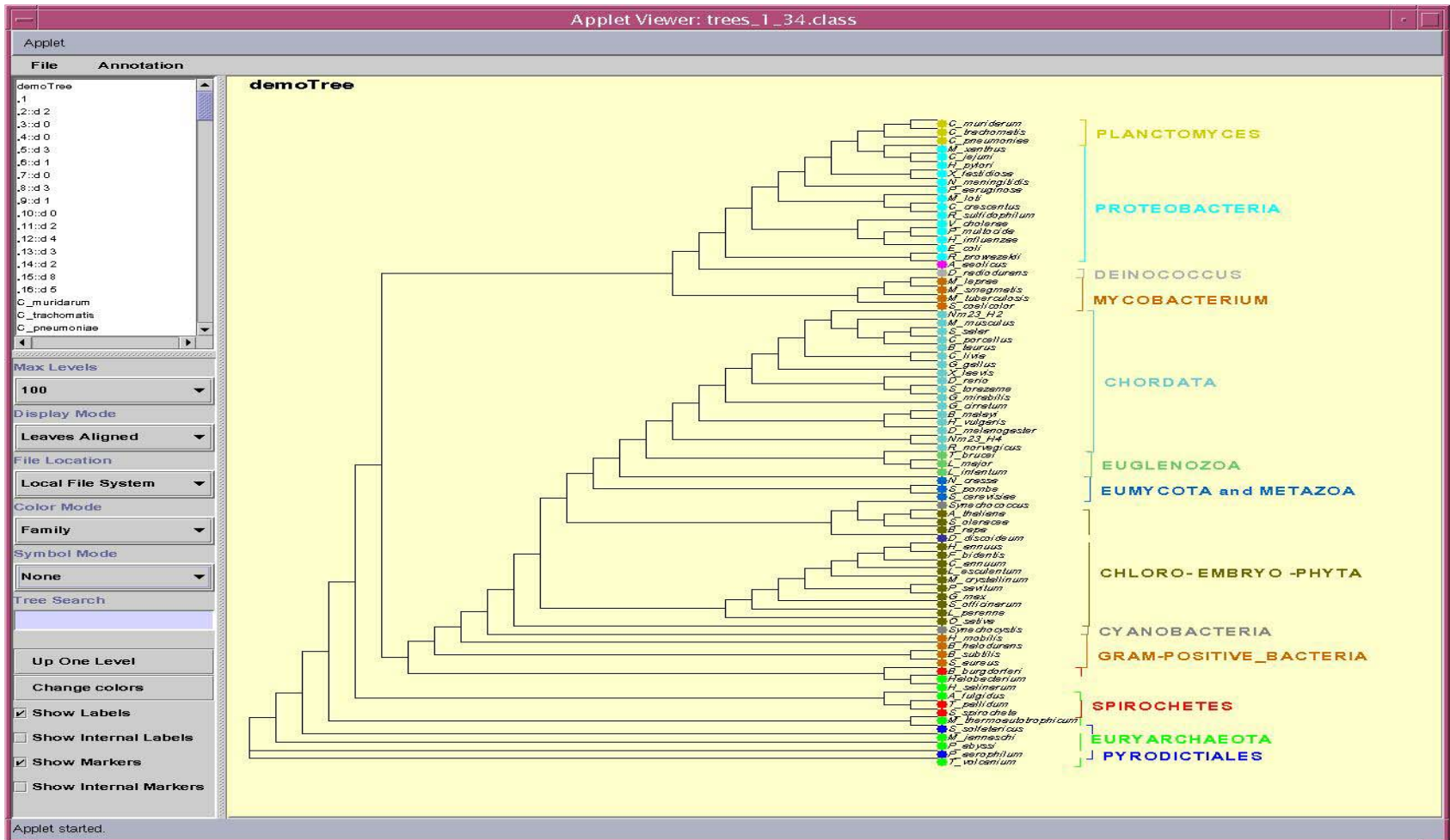
Engineering systems biology

Resource allocation tools

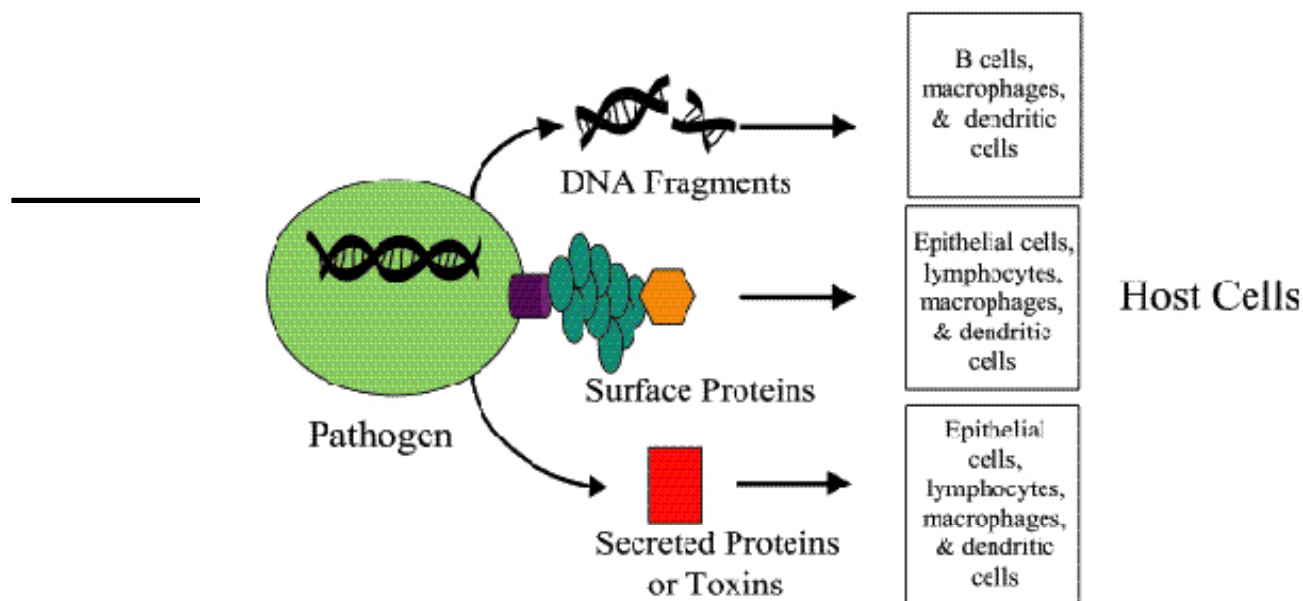
New comparative viewer: *Yersinia pestis* CO-92 plasmid pCD1, Kim plasmid pCD1, *Yersinia enterocolitica* plasmid pYVe227 serotype 9



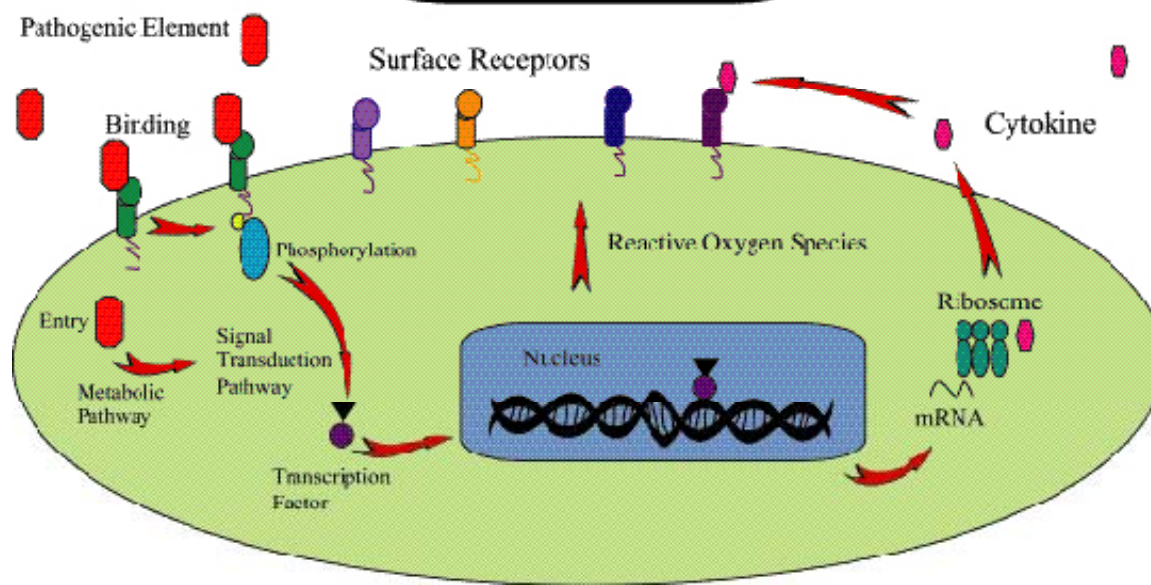
Phylogenetic tree tool: Sample tree showing new annotation feature for grouping terminal taxa



Host-Pathogen Interactions: the Pathogenic Elements



Host Response



In 1962, Nobel Laureate Sir McFarland Burnet:

“One can think of the middle of the 20th century as the end of one of the most important social revolutions in history – the virtual elimination of the infectious disease as a significant factor in social life.”

Acknowledgements

Thomas Brettin
Cathy Cleland
Jason Gans
Gerry Myers
Jian Song
Charlie Strauss
Scott White
Gary Xie
Yan Xu

Russ Altman (Stanford)
Karla Atkins (LANL)
Stephen Eubank (LANL)
Wu Feng (LANL)
Goutam Gupta (LANL)
Lynette Hirschman (MITRE)
Tom Slezak (LLNL)
Gary Strong (NSF)